DOCUMENT RESUME

TE 000 510

ED 022 751

By-Madaus, George F.; Rippey, Robert M.

ZEROING IN ON THE STEP WRITING TEST: WHAT DOES IT TELL A TEACHER?

Pub Date 66

Note-7p.; Paper read at the Annual Meeting of the National Council on Measurement in Education (Chicago, Ill., February 1966).

Available from-National Council of Measurement in Education, Office of Evaluation Services, Michigan State Univ., East Lansing, Mich. (Single copy $2.50).

Journal Cit-Journal of Educational Measurement; v3 n1 p19-25 Spring 1966

EDRS Price MF-$0.25 HC Not Available from EDRS.

Descriptors-COMMUNICATION (THOUGHT TRANSFER), *COMPOSITION SKILLS (LITERARY), ENGLISH INSTRUCTION, EVALUATION CRITERIA, EVALUATION TECHNIQUES, LANGUAGE TESTS, LANGUAGE USAGE, MULTIPLE CHOICE TESTS, PARAGRAPH COMPOSITION, PUNCTUATION, *STUDENT EVALUATION, STUDENT TESTING, TEST INTERPRETATION, *TEST VALIDITY, VERBAL TESTS, *WRITING SKILLS

Identifiers-*Sequential Tests of Educational Progress Writing T, STEP Writing Test

The validity of the multiple-choice Sequential Tests of Educational Progress (STEP) Writing Test (1957) was tested by the University of Chicago Center for the Cooperative Study of Instruction. Seven criteria developed by the center to score essay assignments were used to determine the relationship between STEP and actual writing behavior. Of the four objectives of the STEP test which appeared congruent with four of the essay-grading criteria, a comparison of scores showed a small significant correlation between STEP and the "Punctuation" score, a moderate significant correlation between STEP and the "Usage" and "Effective Organization of the Paragraph" scores, and no significant correlation with the "Sense of Audience and Purpose" score. The analysis indicated that (1) the ability to produce good writing appears only moderately related to the ability to manipulate previously given material on STEP, and (2) the total writing score from STEP does not relate strongly to any of the individual essay-evaluating criteria but does agree moderately with their combined score. The results must be qualified, however, since no measures of parallel form or score reliability were available. (LH)

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY Natl Council of Measurement in Educ. TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U. S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 3, NO. 1
SPRING, 1966

ZEROING IN ON THE STEP WRITING TEST: WHAT DOES IT TELL A TEACHER?[1]

GEORGE F. MADAUS AND ROBERT M. RIPPEY
Center for the Cooperative Study of Instruction
The University of Chicago

A large number of schools administer the STEP Writing Test (1957). Although the publisher of STEP has clearly stated the objectives which STEP was designed to measure, the question frequently arises from classroom teachers, just what can you tell about a student's actual writing behavior from the results of a multiple choice test.

The manual for interpreting scores claims that STEP measures ability to think crtically in writing, to organize materials, to write material appropriate for a given purpose, to write effectively, and to observe conventional usage in punctuation and grammar (p. 7). The manual further states that, "The STEP Writing tests seek to measure comprehensively the full range of skills involved in the process of good writing."

Items on the STEP are classified according to five categories: 1. organization, 2. conventions, 3. critical thinking, 4. effectiveness, 5. appropriateness.

Black (Buros, 1959, p. 593-4) asserts that STEP fails in all but the second category and is only partially successful in measuring conventions. He arrived at his conclusions from his analysis of item content. Perhaps revealing his own biases concerning multiple choice writing tests he concludes "any educator who wishes to measure the full range of skills involved in the process of good writing will resort to writing itself."

Hieronymous (Buros, 1959, p. 595) concludes, again from his analysis of item content, that STEP measures "very effectively higher-order writing skills, particularly those of effectiveness and appropriateness.

The publishers report correlations between STEP Writing and English grades. The difficulty in using such grades as criteria of writing ability lies in the large number of heterogeneous activities subsumed under an English grade. This has been pointed out by the Report of the Commission on English (College Entrance Examination Board, 1965).

Allen, (Buros, 1959, p. 596-7) disturbed by the STEP Writing item content and lack of statistical evidence of validity urges that STEP Writing be compared with other measures of actual writing.

[1]Paper read at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, February, 1966.

ED022751
TE000 518

This paper will describe an exploratory validity study of this type.

During the past two years, the Center for the Cooperative Study of Instruction has been conducting an experiment in the teaching of writing. This experiment has resulted in the evaluation of students using both the STEP Writing Test Level I and a set of seven criterion referenced scales, developed for the purpose of helping graders identify the incidence of specific writing behaviors on the part of students (Rippey, 1965).

These scales were developed by the experienced high school English teachers participating in the project. They first chose seven skills they thought minimally necessary for good writing and then listed behaviors they would accept as evidence of these characteristics. The seven were (1) Punctuation, (2) Usage, (3) Sense of Audience and Purpose, (4) Organization, (5) Use of Detail, (6) Attitude Toward Writing, and (7) A Good Topic and Concluding Sentence. The scales with their behaviors are found in Appendix 1.

The students were given four writing assignments at the beginning of the experiment and the same four assignments seven months later.

Each assignment was structured to evoke samples of writing which then could be graded according to one or more of the seven criterion scales.

The first assignment, *Why Read* was given after all the students had read passages on the importance of reading by Milton, Thoreau, Cervantes, Wolfe, Salinger, and Faulkner. They were directed to present a convincing argument about the importance of reading to other members of the class. This assignment was scored using the Punctuation, Usage, Sense of Audience and Purpose, and Effective Organization of the Paragraph scales.

A paragraph on "the most interesting person I have known" was scored using the Use of Detail Criteria. A paragraph on "everything you have written in the past year" was scored using the Attitude Toward Writing criteria. A final paragraph on any topic the student wished was scored using the Good Topic and Concluding Sentence criteria.

Thus seven scores were obtained from the structured writing assignments on a set of subordinate behaviors, which judiciously applied should result in improved writing on the part of the student.

Four of the stated objectives of STEP Writing appear to be congurent with four of the criterion scales used to score the writing assignments, namely Punctuation, Usage, Sense of Audience and Purpose, and Effective Organization of the Paragraph. We therefore felt that the validity of these four STEP Writing objectives should be tested using the essay scores as criteria.

Procedure and Results

Carefully trained independent readers applied the rating scales to the papers written by a random sample of one-third of the total number of students participating in the experiment.

The freshman, representing the entire experimental population of one school, along with any students who did not have a complete set of scores, were removed for purposes of this study. The remaining group of 101 sophomores, juniors, and seniors drawn from two schools were used in the analyses which follow. Only the post-test data from the experiment were used.

Correlations among the scores assigned by two graders on each of the seven variables are contained in Table 1.

No measures of parallel form or score reliability were available. The table also shows the inter correlation among the criterion derived essay scores.

TABLE 1

Intercorrelations Between The Seven Post-test Essay Scores on the Criterion Reference Scales. (Inter-rater Correlations Appear in Diagonal)

Test	1	2	3	4	5	6	7
1. Punctuation	.72						
2. Usage	.37	.80					
3. Sense of Audience & purpose	—.18	.04	.72				
4. Organization	.05	.23	.37	.76			
5. Use of Detail	—.09	.10	.18	.22	.70		
6. Attitude	.19	.17	.15	.33	.32	.72	
7. Topic & Concluding	.18	.06	—.08	.16	.40	.28	.88
Criterion Score	.23	.42	.16	.32	.05	.40	.25

Eight of the 21 correlations are significantly different from zero at the .05 level or beyond. None of these significant correlations exceed .40.

The administration of the STEP Writing Test in the spring did not adhere to standardization procedures in that 55 instead of 70 actual testing minutes were allowed. However, the mean converted score for all students was 289 which represents a school mean percentile of 96 for grade 11 and 63 for grade 12 according to the STEP school mean norms for fall testing. These figures suggest that the test was neither too difficult nor the time alloted too short for our sample. This time differential must be borne in mind, however, when interpreting our results.

To determine the maximum possible variance shared by STEP and the seven critrion derived essay scores, a multiple regression was performed. This resulted in an R of .58. Thus 34% of the variance of STEP is accounted for by the seven essay scores.

The validity of the STEP objectives for this sample using the essay scores as criteria can be judged from the final row of coefficients in Table 1.

There is a statisticaly significant but small correlation of .23 between STEP and the punctuation score. STEP does not have a statistically significant relation-

ship with our Sense of Audience and Purpose variable. STEP has statistically significant but moderate correlations of .42 and .32 between Usage, and Ability to Organize as measured by our criterion scales.

Discussion

From the above analyses several points seem to emerge. First, abilities to actually produce the component behaviors related to good writing are at best, for this sample, only moderately related to the ability to revise, rearrange, judge, or choose previously given information as on the STEP. The reason for lack of stronger agreement may be due to something closely akin to what the Report of the Commission on English points out. (College Entrance Examination Board, 1965, p. 80).

> "It is not just that analysis is different from synthesis, or that learning how to see and understand is different from learning how to show and to communicate. The difference goes deeper, to the very quick of the student's life, where, like any writer, he exposes himself to public scrutiny, lays his mind bare or all to see."

Secondly, the total writing score of the STEP does not seem to be related strongly to any of the individually important writing behaviors our English teachers felt minimally necessary for good composition. When the seven essay scores are used together they agree moderately well with this writing score given by the STEP.

An alternative interpretation of these data also suggests itself. Since the score or parallel form reliability is unknown and since it is probably less than inter-rater reliability (Gulliksen, p. 212) the actual relationship between STEP Writing and actual writing behavior may be higher than the relationships obtained in the above analyses (Coffman, 1966). If so, these obtained correlations may be suggestive of a higher relationship between STEP Writing and actual writing behavior than can be inferred from the relationships obtained.

For this sample the two essay scores which correlated most highly with the STEP score were the Usage and Attitude Toward Writing scores. The moderate correlation of .40 between STEP and the Attitude Toward Writing variable is interesting. The attitude index gives high scores to those students who report, on their own initiative, a large amount of self-directed, broadly oriented writing which turned out to be pleasant and useful to them. This is, of course, a good description of the student who has been successful in writing up to this time.

We suspect that if teachers were to rank students in order of their writing ability that the rankings would correlate more highly with STEP than with our rating scales. Our scales do not claim to measure global writing ability or to be a comprehensive list of necessary ingredients of good writing; style, fictional and narrative techniques, spelling, and substance are obviously missing.

This study did not attempt to validate this global aspect of STEP but only attempted to see how STEP compared with seven limited aspects of actual writing behavior.

STEP gives a teacher information on how a student stands in relation to a norm group on some kind of global writing score. It did not tell our teachers or students specifically what the examinees could or could not do in writing a composition. It did not offer specific directions for improving writing. This is the conclusion we must draw from these data.

We would recommend that STEP writing be modified to give in addition to an overall score, sub-scores on the various writing behaviors STEP claims it is measuring. We would further suggest that these scores be validated against the actual writing behaviors and not against composite writing performances.

Furthermore, if possible, the scores should be referenced to actual writing behavior as well as to a norm group. This may well involve two separate scores with different interpretations and implications.

References

BUROS, O. K. *Sixth mental measurement Yearbook,* Highland Park, New Jersey: The Gryphon Press, 1959.

COFFMAN, W. E. On the validity of essay tests of achievement, *Journal of Educational Measurement* (in press, 1966).

COLLEGE ENTRANCE EXAMINATION BOARD. *Freedom and Discipline in English,* Report of the Commission on English, New York: College Entrance Examination Board, 1965.

EDUCATIONAL TESTING SERVICE. *Sequential tests of educational progress manual for interpreting scores: Writing.* Princeton: Educational Testing Service, 1957.

GULLIKSEN, H. *Theory of mental tests,* New York: John Wiley and Sons, 1950.

RIPPEY, R. M., *A criterion referenced test in English composition,* Chicago: The Center for the Cooperative Study of Instruction, The University of Chicago, 1965. (Mimeo.)

JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 3, NO. 1
SPRING, 1966

APPENDIX 1
ENGLISH COMPOSITIONS
GRADING SCALES

Variable 1, Punctuation.

Suggested Point Value

1 Does the subject use capital letters correctly?
1 Does the subject terminate his sentences properly?
1 Does the subject employ possessives or contractions, and are they punctuated correctly?
1 Are commas used correctly?
1 Are colons or semicolons used correctly?
1 Does the student employ quotations, and use the quotation marks correctly?
1 If the student uses quotations, are the commas, question marks and periods associated with the quotations placed correctly?
1 If the subject makes no more than a single error in any of the above categories, add 1 additional point.
1 Does the subject use italics, ellipses, exclamation marks, special indentations, or other miscellaneous punctuations correctly?

Variable 2. Usage

Suggested Point Value

1 Does the subject use incomplete sentences?
1 Does the subject use run-on sentences?
1 Does the subject make errors in agreement of subject and verb?
1 Does the subject make errors of pronoun reference?
1 Does the subject misuse any words?
1 Does the subject write any meaningless sentences?
1 Does the subject write any sentences which are obviously awkward?
1 If the subject made no more than a single error in any of the above categories give one extra point.
1 Is the paper free from any miscellaneous errors of usage not covered by the first seven categories?

Variable 3. Sense of Audience and Purpose: Sense of purpose and sense of audience are related. In judging purpose, the following questions might be asked:

Suggested Point Values

1 Can you, the reader, state the purpose of the author?
1 Was it difficult for you to ascertain just what the purpose was?
1 Did the writing deal consistently with a single purpose?
1 Did the writing contribute to the purpose you identified?
1 Would the writing be likely to move the intended audience in the direction intended by the author?
1 Was the language and vocabulary suited for the target audience?
1 Could you identify the audience from reading the paper, or was the paper written in a bland "teacher-pleasing" style?

1 Did the paper show evidence of the author's having thought about the viewpoints of the reader and his biases?
1 Did the author appear to consider those areas where the reader might have difficulty in accepting his argument?

Variable 4. Effective organization of paragraph.

Suggested Point Values

1 Does the author use more than just the simple sentence?
1 Does the author use both compound and complex sentences?
1 Does the author use sentences of varying length? (Are some sentences at least three times as long as others?)
1 Does the author use both positive and negative examples?
2 Does the author arrange his sentences in a logical order such as concrete to abstract, simple to complex, familiar to unfamiliar, geographically, chronologically, etc?
2 Does the order and organization of the sentences serve the purpose of the paper and help to make it more interesting or easier to understand?
1 General effect of the paper as a unified whole.

Variable 5. Selection of Detail to support the purpose of the paragraph.

Suggested Point Values

2 Does the writer use details, or is his paper a jumble of abstractions?
1 Does the writer use both concrete and specific details?
2 Are the details relevant to the purpose?
2 Are the details well chosen and vivid?
2 Are both physical and psychological details included?

Variable 6. Attitude toward writing.

Suggested Point Values

1 Does the writer indicate that he has done much writing?
1 Does the writer express favorable attitudes toward writing, or does he suggest that it is a waste of time?
1 Does the writer show evidence of having enjoyed the writing which he has done?
2 Does the writer indicate that writing has served some useful purpose for him?
2 Has the writer written broadly, or are his writings narrow in scope and purpose?
1 Has the writer written largely in response to assignments or has he done writing on his own?
1 Has the writer chosen serious topics to write about, or trivial ones? Paragraphs two and three might add to this.

Variable 7. Good topic and concluding sentence. (5 points on each)

Suggested Point Value

2 Can you clearly identify a topic and a concluding sentence?
2 Is the placement of these sentences appropriate?
2 Do these sentences have an appropriate impact or effect on the reader?
1 Dos the topic sentence specify the object of the writing and an attitude about it?
2 Do the topic and concluding sentences serve a real purpose, or do you feel that you could do without them?